1/5/3 **(Item 3 from file: 351)**
DIALOG(R)File 351:Derwent WPI

0012868245 - Drawing available
WPI ACC NO: 2002-727258/
XRPX Acc No: N2002-573577
Parallel computer system detects position of node in network based on data transmitted by address conversion entry, when address transmission entry corresponding to transmitted data is selected and registered in cache
Patent Assignee: HITACHI LTD (HITA)
Inventor: KATO S; KAWAMURA T
Patent Family (1 patents, 1 countries)

Patent Number	Kind	Date	Application Number	Kind	Date	Update
JP 2002259213	A	20020913	JP 200153468	A	20010228	200279 B

Priority Applications (no., kind, date): JP 200153468 A 20010228

Patent Details

Number	Kind	Lan	Pg	Dwg	Filing Notes
JP 2002259213	A	JA	12	9	

Alerting Abstract JP A
NOVELTY - The data is transmitted between the nodes (102), through the network (101) by converting the address. The position of a node in the network is detected based on the data transmitted by an address conversion entry, when address conversion entry corresponding to the data to be transmitted is selected from address conversion table and registered with the address conversion table cache.
USE - Parallel computer system having several computers connected through network.
ADVANTAGE - The utilization effectiveness of cache is increased and hence speed of converting the address by cache is increased.
DESCRIPTION OF DRAWINGS - The figure shows the block diagram of hardware incorporated in address conversion function and packet division function.
(Drawing includes non-English language text).
101 Network
102 Node

Title Terms/Index Terms/Additional Words: PARALLEL; COMPUTER; SYSTEM; DETECT; POSITION; NODE; NETWORK; BASED; DATA; TRANSMIT; ADDRESS; CONVERT; ENTER; TRANSMISSION; CORRESPOND; SELECT; REGISTER; CACHE

Class Codes
International Classification (Main): G06F-012/12
(Additional/Secondary): G06F-012/08, G06F-012/10, G06F-015/16, G06F-015/163

File Segment: EPI;
DWPI Class: T01
Manual Codes (EPI/S-X): T01-H03A

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－259213
（P2002－259213A）
(43)公開日　平成14年９月13日(2002.9.13)

(51)Int.Cl.'		識別記号	FI			テーマコード'(参考)
Ｇ０６Ｆ	12/12	５５５	Ｇ０６Ｆ	12/12	５５５	５Ｂ００５
		５０１			５０１	５Ｂ０４５
	12/08	５０７		12/08	５０７Ｇ	
	12/10	５０１		12/10	５０１Ｆ	
	15/16	６４５		15/16	６４５	

審査請求　未請求　請求項の数２　ＯＬ　（全 12 頁）　最終頁に続く

(21)出願番号　特願2001－53468(P2001－53468)

(22)出願日　平成13年２月28日(2001.2.28)

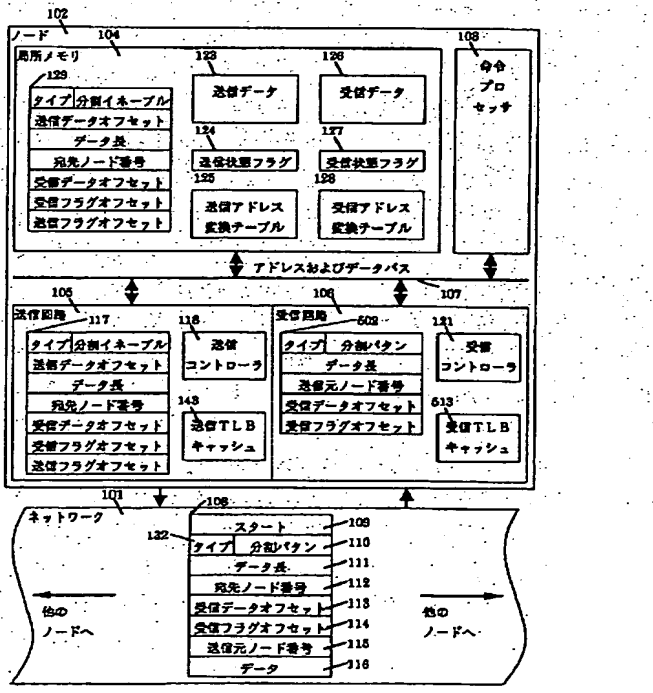
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　加藤　信一
神奈川県秦野市堀山下１番地　株式会社日立製作所エンタープライズサーバ事業部内
(72)発明者　川村　敏雄
神奈川県秦野市堀山下１番地　株式会社日立製作所エンタープライズサーバ事業部内
(74)代理人　100075096
弁理士　作田　康夫

Ｆターム(参考)　5B005 JJ13 KK02 LL11 MM51 NN46 QQ04
5B045 DD12 GG11

(54)【発明の名称】　並列コンピュータシステム

(57)【要約】
【課題】複数のノード間でアドレス変換を行ないながらデータをやり取りするシステムにおいて、複数ページに渡る長いデータを多数転送していると、ＴＬＢキャッシュが溢れる可能性が高くなり、ＴＬＢキャッシュによるアドレス変換の高速化が期待できなくなる。
【解決手段】ＴＬＢキャッシュの各ページエントリに上書き可能であるかないかを示すフィールドを用意し、ページエントリ登録の際に転送するデータの途中のページエントリであればすぐに上書き可能に、またデータの最初か最後であればすぐに上書きできないように優先順位を付ける手段を設ける。

図　1



【特許請求の範囲】

【請求項１】　ネットワークにより接続された複数の処理ノードから構成され、アドレス変換を行ないながらネットワークを介してノード間でデータを転送するコンピュータシステムであって、転送するデータに対応したアドレス変換エントリをアドレス変換テーブルより選択しアドレス変換テーブルキャッシュに登録する際に、アドレス変換エントリの転送するデータに対応する位置を判別可能なことを特徴とする並列コンピュータシステム。

【請求項２】　ネットワークにより接続された複数の処理ノードから構成され、アドレス変換を行ないながらネットワークを介してノード間でデータを転送するコンピュータシステムであって、転送するデータに対応したアドレス変換エントリをアドレス変換テーブルより選択しアドレス変換テーブルキャッシュに登録する際に、アドレス変換エントリの転送するデータに対応する位置によって前記キャッシュ内での優先付けが可能な情報を前記キャッシュのエントリに有することを特徴とする並列コンピュータシステム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、ネットワークによって接続された複数のコンピュータから成るコンピュータシステムに関し、特に多くのコンピュータがローカル及び広域ネットワークを介してパケットを送受信する際にアドレス変換を行なうコンピュータシステムに関係する。

【０００２】

【従来の技術】複数の計算ノード間のメッセージ通信手段として特開平７－２６２１５２号公報の「コンピュータシステム」がある。

【０００３】この公知例では送信データアドレス、送信フラグアドレス、受信データアドレス、受信フラグアドレスは局所メモリの実アドレスを示しており、局所メモリ内の連続領域のデータ送受信が可能である。この公知例のようなシステムに対して、局所メモリの実アドレス空間よりも大きなアドレス空間、即ち、仮想アドレス空間でデータの送受信を行なおうとすると、送信回路および受信回路内にアドレス変換機構を設ける必要がある。アドレス変換機構を設けた送信回路や受信回路がデータ送受信時にアドレス変換する際、仮想アドレスと実アドレスとの関係を示しているアドレス変換テーブルにアクセスする必要がある。しかし、仮想アドレスと実アドレスとの関係をきめ細かく行なおうとすると、アドレス変換テーブルのハードウエア物量としては大きくなり、ハードウエアロジックで実現するのは困難である。そこで、局所メモリに設けることで容易に実現可能になるが、アドレス変換テーブルへのアクセス時間が長くなり、アドレス変換のオーバヘッドが大きくなる。そこで、限られたハードウエアの物量を投入し、アドレス変換テーブルの一部を格納するＴＬＢキャッシュを備えることでアドレス変換テーブルへのアクセス時間を短くなるように工夫している。

【０００４】

【発明が解決しようとする課題】一般的なＴＬＢキャッシュの制御方法としてＦＩＦＯ方式やＬＲＵ方式があり、これらの方式を利用してＴＬＢキャッシュを管理する。図９に仮想アドレス空間と実アドレス空間上のデータの例を示す。ここでは転送するデータ７０１が仮想アドレス空間７０２と実アドレス空間７０３との関係を示すアドレス変換単位をページ７０４と呼ぶ。ページ番号と実アドレス空間の先頭アドレスとの関係を記述したエントリを複数備えたアドレス変換テーブルにより、仮想アドレス空間７０２のページ７０４はこのように自由に実アドレス空間７０３に割り当てることが可能である。この例ではデータ７０１はページ番号１から４までの広範囲に渡っており、実アドレスに変換する際に、ＴＬＢキャッシュに対して４ページ分ものページエントリの登録が必要になる。ＴＬＢキャッシュには限られた数のページエントリしか格納できないので、ページ７０４を複数に渡るような長いデータ７０１を多数転送していると、直ぐに一杯になってしまい溢れる。ページ番号１や４などのデータ転送にはまだ使用されていない領域があって直ぐに再利用される可能性の高いページエントリと、ページ番号２や３などのデータ転送に全て使用されたページエントリは分別なく扱われるため、再利用される前に別のデータ転送によって溢れる可能性が高くなる。ＴＬＢキャッシュのページエントリを再利用する前に溢れて消えてしまうと、再び局所メモリにアクセスしなければならず、ＴＬＢキャッシュによるアドレス変換の高速化が期待できなくなる。

【０００５】

【課題を解決するための手段】上記目的を達成するために、ＴＬＢキャッシュ内の全てのエントリに、上書きを優先してもよいかどうかを示すオーバーライトフィールドを設ける。次に、局所メモリにあるアドレス変換テーブルのエントリをＴＬＢキャッシュに登録するときに、オーバーライトフィールドが有効であるエントリがあればそのエントリに上書きする手段を設ける。さらに、ページエントリをＴＬＢキャッシュに登録後、そのページエントリがデータの最初のエントリであるか途中のエントリであるか、あるいは最後のエントリであるかを判別する手段を設けて、データの途中のページエントリであるときはオーバーライトフィールドを有効にし、データの最初か最後のエントリであればオーバーライトフィールドを無効にする手段を設ける。このようにオーバーライトフィールドを設けることで、直ちに再利用される可能性の低いページエントリから先に後続のデータのページエントリに上書きされ、再利用される可能性の高いページエントリを優先してＴＬＢキャッシュ内に残すこと

ができるので、複数ページにまたがる長いデータを多数転送するシステムにおいて有効である。

【０００６】

【発明の実施の形態】本実施例を図１に示す。これはネットワーク１０１で接続した複数のノードからなる。便宜上、ノード１０２の内の１個のみを図１に示す。図示のノード１０２は送信ノードと受信ノードの両方を示すために用いられているが、実際にはこれらのノードは別のものである。各ノード１０２は命令プロセッサ１０３、適当な容量を備えた局所メモリ１０４、送信回路１０５および受信回路１０６を含む。アドレスおよびデータバス１０７は送信回路１０５にコマンドを出し、命令プロセッサ１０３と局所メモリ１０４の間にデータを転送し、そして局所メモリ１０４と送信回路１０５および受信回路１０６との間でデータを転送するために用いられる。これらの４つの接続はアドレスおよびデータバス１０７で実現されているが、それぞれ独立のアドレス／データ線でもよい。

【０００７】ネットワーク１０１はノード１０２内の送信回路１０５からパケット１０８ｄを受け取り、各パケット１０８に含まれる宛先ノード番号１１２の内容に従って他のノード１０２の受信回路１０６にそれらを送る。本例ではネットワーク１０１はパケット１０８を送信順に受信回路１０６に転送する。

【０００８】図２は送信回路１０５のブロック図である。送信シーケンサ１３０は送信回路１０５全体の制御を行なう。送信コマンドは送信制御ワードブロック１２９（以下、ＴＣＷＢと呼ぶ）の形をとり、これが命令プロセッサ１０３により局所メモリ１０４に記憶される。

【０００９】ＴＣＷＢアドレスレジスタ１３１は命令プロセッサ１０３により書き込み可能なレジスタである。ＴＣＷＢバッファ１１７はＴＣＷＢ１２９のコピーを一時的に保持するために用いられるレジスタファイルである。これはＴＣＷＢと同じフォーマットを有し、そしてパケットタイプ１３２、パケット分割イネーブル１３３、送信データオフセット１３４、データ長１１１、宛先ノード番号１１２、受信データオフセット１１３、受信フラグオフセット１１４、送信フラグオフセット１３５を含む。

【００１０】送信ＦＩＦＯデータバッファ１３６はファーストイン・ファーストアウト（ＦＩＦＯ）データバッファであって、局所メモリ１０４とネットワーク１０１間のデータの緩衝に用いられる。送信直接メモリアクセスコントローラ１３７（以下、送信ＤＭＡＣと呼ぶ）は送信状態フラグ１２４を局所メモリ１０４に書き込むことができ、あるいは局所メモリ１０４からのブロック読み出しを行なうことのできる回路である。このブロック読み出し動作のために、送信ＤＭＡＣ１３７はそのブロックの長さおよびそのブロックのスタートアドレスに初期化される長さカウンタとアドレスカウンタを含む。

【００１１】送信ＤＭＡＣ１３７へのコマンドは送信アドレス変換回路１３８と送信シーケンサ１３０から得られる。ブロック読み出し動作については、アドレスおよびデータバス１０７に出力されるアドレスカウンタの現在の値を用いて局所メモリ読み出し要求が発生され、アドレスカウンタが増分され長さが０となるまで長さカウンタから減分される。

【００１２】局所メモリ１０４から読み出されたデータは送信シーケンサ１３０の制御によりＴＣＷＢバッファ１１７または送信ＦＩＦＯデータバッファ１３６に記憶される。送信ノードレ番号レジスタ１３９は各ノード１０２に固有の、そのノードを識別する番号を含む。

【００１３】ページサイズレジスタ１４０は仮想アドレス空間を複数のページで分割したときの１ページあたりのページサイズを示している。ＴＬＢアドレスレジスタ１４１は局所メモリ１０４にある送信アドレス変換テーブル１２５の先頭アドレスを示している。

【００１４】パケット組み立て部１４２はＴＣＷＢバッファ１１７を送信ノード番号レジスタ１３９から情報を抽出してパケットヘッダを構成し、送信ＦＩＦＯデータバッファ１３６からデータを取り出してパケット１０８を形成し、そして完全なパケット１０８をネットワーク１０１に送り出す回路である。

【００１５】ノード１０２について１つ以上のパケット１０８を1つ以上の他のノード１０２に送るために、送信ノード１０２内の命令プロセッサ１０３は、まず転送されるべき各パケット１０８についての情報として、局所メモリ１０４にＴＣＷＢ１２９を作る。

【００１６】この命令プロセッサ１０４はＴＣＷＢ１２９内に含まれる各フィールドを初期化する。タイプ１３２は初期化されたパケットタイプを示す。本例では全てのパケットタイプは同じであるからタイプ１３２は固定値である。

【００１７】分割イネーブル１３３は１つのパケットで送信するデータ長を制限するかどうかを指定している。送信データオフセット１３４は仮想アドレス空間上の送信データの先頭アドレスを示しており、局所メモリ１０４から送信データを読み出すときには実アドレスに変換する必要があるが、これについては後述する。データ長１１１は転送されるべきデータの長さを示す。宛先ノード番号１１２はそのパケットについての宛先ノードを識別する。受信データオフセット１１３は仮想アドレス空間上の受信データの先頭アドレスを示しており、局所メモリ１０４に受信データを書き込むときには実アドレスに変換する必要があるが、これについては後述する。受信フラグオフセット１１４は宛先ノードの局所メモリ１０４に、受信状態フラグ１２７をポイントすべきところを示すが、ポイントするには実アドレスに変換する必要があり、これについては後述する。

【００１８】パケット送信プロセスをスタートするため

に、命令プロセッサ１０４はアドレスおよびデータバス１０７を用いてＴＣＷＢ１２９の先頭アドレスをＴＣＷＢアドレスレジスタ１３１に書き込む。送信シーケンサxxxはこのレジスタへの書き込みにより動作を開始する。

【００１９】まず、送信シーケンサ１３０はＴＣＷＢアドレスレジスタ１３１に書き込まれたアドレスを用いて送信ＤＭＡＣ１３７にＴＣＷＢ１２９の読み出しを命令する。ＴＣＷＢ１２９はアドレスおよびデータバス１０７を通って局所メモリ１０４から読み出され、そしてＴＣＷＢバッファ１１７に記憶される。次に、送信シーケンサ１３０は送信アドレス変換回路１３８に対して送信データの読み出しを行なうよう命令する際、オフセット情報と送信データ長情報、分割パタン情報の３つを渡す必要がある。しかし、ＴＣＷＢバッファ１２９内の分割イネーブル１３３が有効であるか無効であるかによって値が異なる。

【００２０】分割イネーブル１３３が無効である場合、本例では０’であるとき、送信アドレス変換回路１３８に渡すオフセット情報と送信データ長情報、分割パタン情報はそれぞれ次の通りとなる。

（オフセット情報）＝（ＴＣＷＢバッファ１１７内の送信データオフセット１３４）

（送信データ長情報）＝（データ長１１１）

（分割パタン情報）＝００’

分割イネーブル１３３が有効である場合、本例では１’であるとき、ＴＣＷＢバッファ１１７内のデータ長１１１が分割パケット長（本例では固定値）より大きいか小さいかによって、送信シーケンサ１３０の処理が２つに分けられる。

【００２１】データ長１１１が分割パケット長と同じか小さい場合、分割イネーブル１３３が０’であるときと同様、ＴＣＷＢバッファ１１７内の送信データオフセット１３４とデータ長１１１、分割パタン情報００’を送信アドレス変換回路１３８に渡す。

【００２２】データ長１１１が分割パケット長より大きい場合、パケットのデータ長が分割パケット長より大きくならないように複数のパケットに分けて送信する。そのため、送信シーケンサ１３０は送信アドレス変換回路１３８に対して、複数に分割して送信データ読み出し命令を行なう。

【００２３】分割の最初の送信データの読み出し命令では、オフセット情報と送信データ長情報、分割パタン情報がそれぞれ次の通りとなる。

（オフセット情報）＝（ＴＣＷＢバッファ１１７内の送信データオフセット１３４）

（送信データ長情報）＝分割パケット長

（分割パタン情報）＝１０’

分割途中の送信データの読み出し命令では、オフセット情報と送信データ長情報、分割パタン情報がそれぞれ次の通りとなる。

（オフセット情報）＝（ＴＣＷＢバッファ１１７内の送信データオフセット１３４）＋（（分割パケット長）×（（分割開始からの順番）－１））

（送信データ長情報）＝分割パケット長

（分割パタン情報）＝１１’

分割の最後の送信データの読み出し命令では、オフセット情報と送信データ長情報、分割パタン情報がそれぞれ次の通りとなる。

（オフセット情報）＝（ＴＣＷＢバッファ１１７内の送信データオフセット１３４）＋（（分割パケット長）×（（分割開始からの順番）－１））

（送信データ長情報）＝（送信データ長を分割パケット長で割った余り）

（分割パタン情報）＝０１’

図５に仮想アドレス空間と実アドレス空間の関係を示す。仮想アドレス空間１４９は複数のページ１５０に別れていて、ページ単位で実アドレス空間１５１上の連続領域に割り当てられている。ページの大きさはページサイズレジスタ１４０により決まる。なお、送信データ１２３の大きさはページサイズに無関係のため、複数のページ１５０に渡る場合があるが、本例では、送信データ１２３が複数のページ１５０に渡らない場合について記述している。該当するページ番号が判別したら、局所メモリ１０４にある送信アドレス変換テーブル１２５を参照して実アドレスに変換する。送信アドレス変換テーブル１２５は、局所メモリ１０４上の連続アドレス領域にページ番号０、１、２…ｎ番まで順番に並べており、送信アドレス変換テーブル１２５の先頭アドレスはＴＬＢアドレスレジスタ１４１に示されている。仮想アドレス空間上の送信データの先頭アドレスはＴＣＷＢ１２９内の送信データオフセット１３４で示されており、送信データ１２３がどのページに存在しているかが判別可能である。送信アドレス変換テーブルから、該当するページ１５０の先頭実アドレスを求める。さらに、送信データオフセット１３４をページサイズで割った余りがページ内の送信データオフセットであり、このオフセットを先に求めた該当するページ１５０の先頭実アドレスに加算した値が送信データの先頭実アドレスとなる。

【００２４】送信アドレス変換回路１３８は、局所メモリ１０４にある送信データの先頭実アドレスと送信データ長を求めて送信ＤＭＡＣ１３７にデータを読み出す指示と、仮想アドレスから実アドレスへの変換の効率向上のために使用する送信ＴＬＢキャッシュ１４３の管理を行なう。詳細は図３、図４に示す手順の通りに行なう。

【００２５】第１に、送信アドレス変換回路１３８は、仮想アドレス空間上のページ番号を求めるため、オフセット情報をページサイズレジスタの値で割って商を計算する。

【００２６】第２に、送信アドレス変換回路１３８は送

信ＴＬＢキャッシュ１４３に、先に求めたページ番号が存在するかどうかを確認する。送信ＴＬＢキャッシュ１４３は図２に示すように、バリッド１４５、オーバーライト１４６、ページ番号１４７、ページ内先頭実アドレス１４８の４つの要素を１つのエントリ１４４として複数エントリ１４４存在する。送信ＴＬＢキャッシュ１４３の全エントリ１４４を見て、バリッド１４５が有効でかつページ番号が同一のエントリ１４４があるかどうかを探す。エントリ１４４が存在した場合、第１０項の処理へ移る。

【００２７】第３に、同一のエントリ１４４が無い場合、局所メモリ１０４の送信アドレス変換テーブル１２５内にあるエントリの先頭アドレスを求める必要がある。先に求めたページ番号にＴＬＢエントリサイズ、本例では固定値を掛け算し、ＴＬＢアドレスレジスタ１４１の値を加算する。

【００２８】第４に、エントリの先頭アドレスを求めたら、送信ＤＭＡＣ１３７に対してエントリを局所メモリ１０４から読み出すように命令する。

【００２９】第５に、エントリが局所メモリ１０４から読み出されるのを待つ。

【００３０】第６に、送信ＴＬＢキャッシュ１４３にあるエントリ１４４に上書き可能なエントリ１４４があるかどうかを確認する。具体的には、送信ＴＬＢキャッシュ１４３にあるエントリ１４４の中で、オーバーライト１４６が１’のエントリが存在するかどうかを確認する。

【００３１】第７に、送信ＴＬＢキャッシュ１４３にあるエントリ１４４の中で、バリッド１４５が１’でかつオーバーライト１４６が１’であるエントリ１４４が存在する場合、そのエントリ１４４に対して、局所メモリ１０４より読み出されたエントリを上書きする。具体的には、先の計算で求めたページ番号と局所メモリ１０４より読み出されたエントリにある該当ページの先頭実アドレスを書き込む。バリッド１４５が１’でかつオーバーライト１４６が１’であるエントリ１４４が存在しない場合、ＦＩＦＯ方式やＬＲＵ方式などの規則によってＴＬＢエントリを１つ選択する。選択したＴＬＢエントリのバリッド１４５が１’の場合、読み出されたエントリを上書きし、バリッド１４５が０’であれば新たに読み出されたエントリを登録する。エントリの登録とは、前述の上書き時の処理に加えてさらに、バリッド１４５を１’に設定することを指す。バリッド１４５とはエントリ１４４そのものが有効であるか無効であるかを示しており、一般的にＦＩＦＯ方式やＬＲＵ方式などで管理されたエントリ１４４の情報である。

【００３２】第８に、上書きまたは登録したエントリ１４４に対して、上書き設定をどうするかを判定する。上書き設定とは、そのエントリ１４４のオーバーライト１４６を設定することを指す。次の４つの判定条件がある。

条件１　分割パタン情報が００’で送信データの最初のページのエントリである。

条件２　分割パタン情報が００’で送信データの最後のページのエントリである。

条件３　分割パタン情報が１０’で送信データの最初のページのエントリである。

条件４　分割パタン情報が０１’で送信データの最後のページのエントリである。

【００３３】第９に、上記の４つの条件のうち、いずれも成立しない場合、オーバーライト１４６を１’設定する。いずれかが成立した場合、オーバーライト１４６を０’に設定する。上記条件１～２により、分割されないパケットの送信データの最初と最後のページのエントリのオーバーライト１４６を０’に設定することにより、前述の第７項での上書きが行われなくなるが、途中のページのエントリのオーバーライト１４６を１’に設定することにより、第７項での上書きが行われることになる。また、上記条件３～４により、分割された先頭パケットの送信データの最初のページのエントリと、分割された末尾パケットの送信データの最初のページのエントリのオーバーライト１４６を０’に設定することにより、第７項での上書きが行われなくなるが、分割される前の連続した送信データとして着目すると途中のページのエントリのオーバーライト１４６を１’に設定することにより、第７項での上書きが行われることになる。

【００３４】第１０に、第４項で求めたページ番号のＴＬＢエントリ内にある該当ページの先頭実アドレスを使って、送信データの先頭実アドレスを求める。送信データの先頭アドレス情報をページサイズで割り、その余りを該当ページの先頭実アドレスに加算した値が送信データの先頭実アドレスである。

【００３５】このようにして、送信アドレス変換回路１３８は、送信データの先頭実アドレスを求め、送信データ長情報とともに送信ＤＭＡＣ１３７に渡して、データの読み出しを命令する。

【００３６】送信シーケンサは、局所メモリ１０４より読み出された送信データをＦＩＦＯデータバッファ１４１に記憶し、そして次に分割パタン情報をパケット組み立て部１４２に渡して、パケット１０８の送信開始を命令する。

【００３７】パケット組み立て部１４２は、まずパケット１０８のスタートをマークする固定スタートコード１０９を送る。また、パケット組み立て部１４２は、ＴＣＷＢバッファ１１７と送信ノード番号レジスタ１３９、送信シーケンサ１３０から渡される分割パタン情報の３つから、タイプ１３２、分割パタン１１０、宛先ノード番号１１２、受信データオフセット１１３、受信フラグオフセット１１４、送信元ノード番号１１５を取り出してそれを送信する。

【００３８】次に、パケット組み立て部１４２は送信ＦＩＦＯデータバッファ１３６から送信データを取り出し、それを送ってパケッデータ１１６を形成する。なお、送信ＦＩＦＯデータバッファ１３６はこれと並行して送信ＤＭＡＣ１３７により局所メモリ１０４から読み出された送信データを記憶している。

【００３９】送信データを全て送信し終えると、パケット組み立て部１４２は送信シーケンサ１３０にパケット送信処理完了を通知する。送信シーケンサ１３０は、送信アドレス変換回路１３８にＴＣＷＢバッファ１１７に記憶された送信フラグオフセット１３５を渡して、そのパケットが送信されたことを示す値を送信状態フラグ１２４に書き込むことを命令する。本例では送信状態フラグ１２４は、パケット１０８が送信されたか、まだ送信されていないかを意味する２つの可能な値の１つを有することができる。

【００４０】本例では、送信状態フラグ１２４は仮想アドレス空間上のページサイズよりも小さい大きさで、さらに、２つのページをまたぐことが無いように、あらかじめ送信フラグオフセット１３５の値を制限されているので、送信データの読み出し命令のときと同様に、送信状態フラグ１２４が複数のページに渡らない場合について記述する。

【００４１】第１に、送信アドレス変換回路１３８は、仮想空間上のページ番号を求めるため、送信フラグオフセットをページサイズレジスタの値で割って商を計算する。

【００４２】第２に、送信アドレス変換回路１３８は送信ＴＬＢキャッシュ１４３に、先に求めたページ番号が存在するかどうかを確認する。具体的には、送信ＴＬＢキャッシュ１４３の全エントリ１４４を見て、バリッド１４５が有効でページ番号が同一のエントリ１４４があるかどうかを探す。エントリ１４４が存在した場合、第９項の処理へ移る。

【００４３】第３に、同一のエントリが無い場合、局所メモリ１０４にあるエントリアドレスを求めるため、先に求めたページ番号にエントリサイズ、この実施例では固定値を掛け算し、ＴＬＢアドレスレジスタ１４１の値を加算する。

【００４４】第４に、エントリアドレスを求めたら、送信ＤＭＡＣ１３７に対してエントリアドレスの指すエントリを局所メモリ１０４から読み出すように命令する。

【００４５】第５に、エントリが局所メモリ１０４から読み出されるのを待つ。

【００４６】第６に、送信ＴＬＢキャッシュ１４３にあるエントリ１４４に上書き可能なエントリ１４４があるかどうかを確認する。具体的には、送信ＴＬＢキャッシュ１４３にあるエントリ１４４の中で、オーバーライト１４６が1’のエントリが存在するかどうかを確認する。

【００４７】第７に、送信ＴＬＢキャッシュ１４３にあるエントリの中で、バリッド１４５が1’でかつオーバーライト１４６が1’であるエントリが存在する場合、そのエントリ１４４に対して、局所メモリ１０４より読み出されたエントリを上書きする。具体的には、先の計算で求めたページ番号と読み出されたエントリにある該当ページの先頭実アドレスを書き込む。バリッド１４５が1’でかつオーバーライト１４６が1’であるエントリ１４４が存在しない場合、ＦＩＦＯ方式やＬＲＵ方式などの規則によってエントリ１４４を１つ選択する。選択したエントリ１４４のバリッド１４５が1’の場合、読み出されたエントリを上書きし、バリッド１４５が0’であれば新たに読み出されたエントリ１４４を登録する。エントリ１４４の登録とは、前述の上書き時の処理に加えてさらに、バリッド１４５を1’に設定することを指す。

【００４８】第８に、上書きまたは登録したエントリ１４４に対して、上書き設定する。上書き設定とは、そのエントリのオーバーライト１４６を設定することを指す。送信状態フラグ１２４は先に述べた通りページサイズよりも小さく、再利用される可能性があるのでオーバーライト１４６は0’に設定する。

【００４９】第９に、第４項で求めたページ番号のエントリ１４４内にある該当ページの先頭実アドレス１４８を使って、送信状態フラグ１２４の先頭実アドレスを求める。送信フラグオフセット１３５をページサイズで割り、その余りを該当ページの先頭実アドレス１４８に加算した値が送信状態フラグ１２４の先頭実アドレスである。

【００５０】このようにして、送信アドレス変換回路１３８は、送信状態フラグ１２４の先頭実アドレスを求め、送信ＤＭＡＣ１３７に渡して、送信状態フラグ１２４の書き込みを命令する。

【００５１】ネットワーク１０１は送信回路１０５からのパケット１０８を受け、そして宛先ノード番号１１２に従い、宛先ノードの受信回路１０６にそのパケット１０８を転送する。本例のネットワーク１０１は、複数のパケット１０８が１つのノードから他のノードに送られるときは、それらのパケット１０８が送信順に入るように構成されている。

【００５２】図６は受信回路１０６のブロック図である。受信シーケンサ５０１はパケット受信状態フラグ１２７の書き込み処理を含む受信回路１０６全体の制御を行なう。パケットヘッダバッファ５０２はパケットヘッダのコピーを一時的に記憶するために用いられる。このバッファ５０２のフォーマットはパケットヘッダのそれとほとんど同じであり、それはパケットタイプ５０３、分割パタン５０４、データ長５０５、受信データオフセット５０６、受信フラグオフセット５０７、送信元ノード番号５０８を含む。

【００５３】パケット受信部５０９はパケット１０１が到着すると、到着したことを受信シーケンサ５０１に伝えるとともに、パケットヘッダをパケットヘッダバッファ５０２へ、受信データを受信ＦＩＦＯデータバッファ５１０に格納する。受信ＦＩＦＯデータバッファ５１０はネットワーク１０１からのデータを受けて、そのデータが局所メモリ１０４に書き込まれる前に緩衝するバッファである。

【００５４】受信直接メモリアクセスコントローラ５１２（以下、受信ＤＭＡＣと呼ぶ）は受信状態フラグ１２７を局所メモリ１０４に書き込むことができ、あるいは局所メモリ１０４にブロック書き込みを行なうことのできる回路である。このブロック書き込み動作のために、受信ＤＭＡＣ５１２はそのブロックの長さおよびそのブロックのスタートアドレスに初期化される長さカウンタとアドレスカウンタを含む。

【００５５】受信ＤＭＡＣ５１２へのコマンドは受信アドレス変換回路５１１と受信シーケンサ５０１から得られる。ブロック書き込み動作については、アドレスおよびデータバス１０７に出力されるアドレスカウンタの現在の値を用いて局所メモリ書き込み要求が発生され、アドレスカウンタが増分され長さが０となるまで長さカウンタから減分される。

【００５６】以下、受信回路１０６の動作を説明する。受信回路１０６内のパケット受信部５０９はパケット１０８のスタートマーカ１０９を受け取ると、受信シーケンサ５０１にパケットの到着を知らせる。受信したパケットの残りの部分からのマーカは分離して捨てる。パケットヘッダのうち、ヘッダバッファ５０２に該当するフィールドのみをヘッダバッファ５０２に格納し、パケットヘッダのデータ長１１１を用いて受信ＦＩＦＯデータバッファ５１０にデータを格納する。

【００５７】受信シーケンサ５０１は受信アドレス変換回路５１１に対して受信データの書き込みを行なうよう命令する際、オフセット情報と受信データ長情報、分割パタン情報の３つを渡す。３つの情報はそれぞれ次の通りとなる。

（オフセット情報）＝（パケットヘッダバッファ５０２内の受信データオフセット５０６）

（受信データ長情報）＝（受信データ長５０５）

（分割パタン情報）＝（パケットヘッダバッファ５０２内の分割パタン５０４）

受信アドレス変換回路５１１は、局所メモリ１０４にある受信データの先頭実アドレスと受信データ長を求めて受信ＤＭＡＣ５１２にデータを書き込む指示と、仮想アドレスから実アドレスへの変換の効率向上のために使用する受信ＴＬＢキャッシュ５１３の管理を行なう。詳細は図７、図８に示す手順の通りに行なう。

【００５８】第１に、受信アドレス変換回路５１１は、仮想アドレス空間上のページ番号を求めるため、オフセット情報をページサイズレジスタの値で割って商を計算する。

【００５９】第２に、受信アドレス変換回路５１１は受信ＴＬＢキャッシュ５１３に、先に求めたページ番号が存在するかどうかを確認する。受信ＴＬＢキャッシュ５１３は図６に示すように、バリッド５１７、オーバーライト５１８、ページ番号５１９、ページ内先頭実アドレス５２０の４つの要素を１つのエントリ５１６として複数エントリ５１６存在する。受信ＴＬＢキャッシュ５１３の全エントリ５１６を見て、バリッド５１７が有効でかつページ番号が同一のエントリ５１６があるかどうかを探す。エントリ５１６が存在した場合、第１０項の処理へ移る。

【００６０】第３に、同一のエントリ５１６が無い場合、局所メモリ１０４の受信アドレス変換テーブル１２８内にあるエントリの先頭アドレスを求める必要がある。先に求めたページ番号にＴＬＢエントリサイズ、本例では固定値を掛け算し、ＴＬＢアドレスレジスタ５１５の値を加算する。

【００６１】第４に、エントリの先頭アドレスを求めたら、受信ＤＭＡＣ５１２に対してエントリを局所メモリ１０４から読み出すように命令する。

【００６２】第５に、エントリが局所メモリ１０４から読み出されるのを待つ。

【００６３】第６に、受信ＴＬＢキャッシュ５１３にあるエントリ５１６に上書き可能なエントリ５１６があるかどうかを確認する。具体的には、受信ＴＬＢキャッシュ５１３にあるエントリ５１６の中で、オーバーライト５１８が1’のエントリが存在するかどうかを確認する。

【００６４】第７に、受信ＴＬＢキャッシュ５１３にあるエントリ５１６の中で、バリッド５１７が1’でかつオーバーライト５１８が1’であるエントリ５１６が存在する場合、そのエントリ５１６に対して、局所メモリ１０４より読み出されたエントリを上書きする。具体的には、先の計算で求めたページ番号と局所メモリ１０４より読み出されたエントリにある該当ページの先頭実アドレスを書き込む。バリッド５１７が1’でかつオーバーライト５１８が1’であるエントリ５１６が存在しない場合、ＦＩＦＯ方式やＬＲＵ方式などの規則によってＴＬＢエントリを１つ選択する。選択したＴＬＢエントリのバリッド５１７が1’の場合、読み出されたエントリを上書きし、バリッド５１７が0’であれば新たに読み出されたエントリを登録する。エントリの登録とは、前述の上書き時の処理に加えてさらに、バリッド５１７を1’に設定することを指す。バリッド５１７とはエントリ５１６そのものが有効であるか無効であるかを示しており、一般的にＦＩＦＯ方式やＬＲＵ方式などで管理されたエントリ５１６の情報である。

【００６５】第８に、上書きまたは登録したエントリ５

１６に対して、上書き設定をどうするかを判定する。上書き設定とは、そのエントリ５１６のオーバーライト５１８を設定することを指す。次の４つの判定条件がある。

条件１　分割パタン情報が００’で受信データの最初のページのエントリである。

条件２　分割パタン情報が００’で受信データの最後のページのエントリである。

条件３　分割パタン情報が１０’で受信データの最初のページのエントリである。

条件４　分割パタン情報が０１’で受信データの最後のページのエントリである。

【００６６】第９に、上記の４つの条件のうち、いずれも成立しない場合、オーバーライト５１８を１’設定する。いずれかが成立した場合、オーバーライト５１８を０’に設定する。上記条件１～２により、分割されないパケットの受信データの最初と最後のページのエントリのオーバーライト５１８を０’に設定することにより、前述の第７項での上書きが行われなくなるが、途中のページのエントリのオーバーライト５１８を１’に設定することにより、第７項での上書きが行われることになる。また、上記条件３～４により、分割された先頭パケットの受信データの最初のページのエントリと、分割された末尾パケットの受信データの最初のページのエントリのオーバーライト５１８を０’に設定することにより、第７項での上書きが行われなくなるが、分割される前の連続した受信データとして着目すると途中のページのエントリのオーバーライト５１８を１’に設定することにより、第７項での上書きが行われることになる。

【００６７】第１０に、第４項で求めたページ番号のＴＬＢエントリ内にある該当ページの先頭実アドレスを使って、受信データの先頭実アドレスを求める。受信データの先頭アドレス情報をページサイズで割り、その余りを該当ページの先頭実アドレスに加算した値が受信データの先頭実アドレスである。

【００６８】このようにして、受信アドレス変換回路５１１は、受信データの先頭実アドレスを求め、受信データ長情報とともに受信ＤＭＡＣ１５１２に渡して、データの書き込みを命令する。

【００６９】受信ＤＭＡＣ５１２内の長さカウンタが０になると、受信データの全てが局所メモリ１０４に書き込まれたことになる。受信シーケンサ５０１は受信アドレス変換回路５１１にパケットヘッダバッファ５０２に記憶された受信フラグオフセット５０７を渡して、そのパケットが受信されたことを示す値を受信状態フラグ１２７の書き込みを命令する。本例では受信状態フラグ１２７は、パケット１０８が受信されたか、まだ受信されていないかを意味する２つの可能な値の１つを有することができる。本例では、受信状態フラグ１２７は仮想アドレス空間上のページサイズよりも小さい大きさで、さらに、２つのページをまたぐことが無いように、あらかじめ受信フラグオフセット５０７の値を制限されているので、受信データの書き込み命令のときと同様に、受信状態フラグ１２７が複数のページに渡らない場合について記述する。

【００７０】第１に、送信アドレス変換回路１３８は、仮想空間上のページ番号を求めるため、送信フラグオフセットをページサイズレジスタの値で割って商を計算する。

【００７１】第２に、受信アドレス変換回路５１１は受信ＴＬＢキャッシュ５１３に、先に求めたページ番号が存在するかどうかを確認する。具体的には、受信ＴＬＢキャッシュ５１３の全エントリ５１６を見て、バリッド５１７が有効でページ番号が同一のエントリ５１６があるかどうかを探す。エントリ５１６が存在した場合、第９項の処理へ移る。

【００７２】第３に、同一のエントリが無い場合、局所メモリ１０４にあるエントリアドレスを求めるため、先に求めたページ番号にエントリサイズ、この実施例では固定値を掛け算し、ＴＬＢアドレスレジスタ５１５の値を加算する。

【００７３】第４に、エントリアドレスを求めたら、受信ＤＭＡＣ５１２に対してエントリアドレスの指すエントリを局所メモリ１０４から読み出すように命令する。

【００７４】第５に、エントリが局所メモリ１０４から読み出されるのを待つ。

【００７５】第６に、受信ＴＬＢキャッシュ５１３にあるエントリ５１６に上書き可能なエントリ５１６があるかどうかを確認する。具体的には、受信ＴＬＢキャッシュ５１３にあるエントリ５１６の中で、オーバーライト５１７が１’のエントリが存在するかどうかを確認する。

【００７６】第７に、受信ＴＬＢキャッシュ５１３にあるエントリの中で、バリッド５１７が１’でかつオーバーライト５１８が１’であるエントリが存在する場合、そのエントリ５１６に対して、局所メモリ１０４より読み出されたエントリを上書きする。具体的には、先の計算で求めたページ番号と読み出されたエントリにある該当ページの先頭実アドレスを書き込む。バリッド５１７が１’でかつオーバーライト５１８が１’であるエントリ５１６が存在しない場合、ＦＩＦＯ方式やＬＲＵ方式などの規則によってエントリ５１６を１つ選択する。選択したエントリ５１６のバリッド５１７が１’の場合、読み出されたエントリを上書きし、バリッド５１７が０’であれば新たに読み出されたエントリ５１６を登録する。エントリ５１６の登録とは、前述の上書き時の処理に加えてさらに、バリッド５１７を１’に設定することを指す。

【００７７】第８に、上書きまたは登録したエントリ５１６に対して、上書き設定する。上書き設定とは、その

エントリのオーバーライト５１８を設定することを指す。受信状態フラグ１２７は先に述べた通りページサイズよりも小さく、再利用される可能性があるのでオーバーライト５１８は0’に設定する。

【００７８】第９に、第４項で求めたページ番号のエントリ５１６内にある該当ページの先頭実アドレス５２０を使って、受信状態フラグ１２７の先頭実アドレスを求める。受信フラグオフセット５０６をページサイズで割り、その余りを該当ページの先頭実アドレス５２０に加算した値が受信状態フラグ１２７の先頭実アドレスである。

【００７９】このようにして、受信アドレス変換回路５１１は、受信状態フラグ１２７の先頭実アドレスを求め、受信ＤＭＡＣ５１２に渡して、受信状態フラグ１２７の書き込みを命令する。

【００８０】宛先ノードの命令プロセッサ１０３は任意の時点でパケット１０８の受信状態フラグ１２７を読み取るとこにより、そのパケット１０８の受信を確認できる。

【００８１】本発明の実施例は、連続したデータを複数のノード間で転送する際に、利用効率の低いアドレス変換テーブルエントリをＴＬＢキャッシュに常駐させないようにして、ＴＬＢキャッシュの利用効率を上げ、アドレス変換を高速化するための手段を与えるものである。

【００８２】

【発明の効果】本発明の並列コンピュータシステムではＴＬＢキャッシュの各ページエントリに上書き可能であるかないかを示すフィールドを設けて、ページエントリ登録の際に転送するデータの途中のページエントリであればすぐに上書き可能に、またデータの最初か最後であればすぐに上書きできないように優先順位を付けることが可能になる。これにより、データの最初または最後のページに相当する、データ転送にはまだ使用されていない領域があって再利用される可能性の高いページエントリと、データ転送の際に全ての領域を使用されたページエントリを区別して扱えるため、再利用される前に別のデータ転送によって溢れる可能性が低くなりＴＬＢキャッシュの利用効率が上がってアドレス変換の高速化が期待できる。

【図面の簡単な説明】

【図１】アドレス変換機能およびパケット分割機能を組み込んだハードウエアのブロック図。

【図２】本発明を組み込んだ送信回路のブロック図。

【図３】本発明を組み込んだ送信アドレス変換回路における送信データのアドレス変換処理手順を示す図。

【図４】図３のａに続く送信データのアドレス変換処理手順を示す図。

【図５】送信データのアドレスを変換するときの仮想アドレス空間と実アドレス空間およびアドレス変換テーブルとの関係を示した図。

【図６】本発明を組み込んだ受信回路のブロック図。

【図７】本発明を組み込んだ受信アドレス変換回路における受信データのアドレス変換処理手順を示す図。

【図８】図７のａに続く受信データのアドレス変換処理手順を示す図。

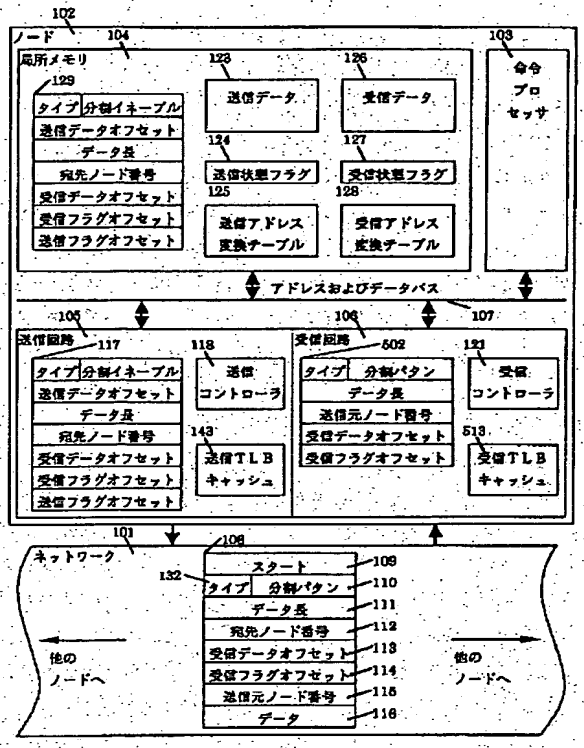
【図９】従来の転送するデータのアドレスを変換するときの仮想アドレス空間と実アドレス空間の関係を示した図。

【符号の説明】

１０１…ネットワーク、１０２…ノード、１０３…命令プロセッサ、１０４…局所メモリ、１０５…送信回路、１０６…受信回路、１０７…アドレスおよびデータバス、１０８…パケット、１０９…スタートコード、１１０…分割パタン、１１１…データ長、１１２…宛先ノード番号、１１３…受信データオフセット、１１４…受信フラグオフセット、１１５…送信先ノード番号、１１６…データ、１１７…ＴＣＷＢバッファ、１１８…送信コントローラ、１２１…受信コントローラ、１２３…送信データ、１２４…送信状態フラグ、１２５…送信アドレス変換テーブル、１２６…受信データ、１２７…受信状態フラグ、１２８…受信アドレス変換テーブル、１２９転送制御ワードブロック（ＴＣＷＢ）、１３１…ＴＣＷＢアドレスレジスタ、１３２…タイプ、１３３…分割イネーブル、１３４…送信データオフセット、１３５…送信フラグオフセット、１３６…送信ＦＩＦＯデータバッファ、１３７…送信ＤＭＡＣ、１３８…送信アドレス変換回路、１３９…送信ノード番号レジスタ、１４０…ページサイズレジスタ、１４１…ＴＬＢアドレスレジスタ、１４２…パケット組み立て部、１４３…送信ＴＬＢキャッシュ、１４４…ページエントリ、１４５…バリッド、１４６…オーバーヘッド、１４７…ページ番号、１４８…ページ内先頭実アドレス、１４９…仮想アドレス空間、１５０…ページ、１５１…実アドレス空間、５０１…受信シーケンサ、５０２…パケットヘッダバッファ、５０３…タイプ、５０４…分割パタン、５０５…データ長、５０６…受信データオフセット、５０７…受信フラグオフセット、５０８…送信元ノード番号、５０９…パケット受信部、５１０…受信ＦＩＦＯデータバッファ、５１１…受信アドレス変換回路、５１２…受信ＤＭＡＣ、５１３…受信ＴＬＢキャッシュ、５１４…ページサイズレジスタ、５１５…ＴＬＢアドレスレジスタ、５１６…ページエントリ、５１７…バリッド、５１８…オーバーライト、５１９…ページ番号、５２０…ページ内先頭実アドレス、７０１…データ、７０２…仮想アドレス空間、７０３…実アドレス空間、７０４…ページ。

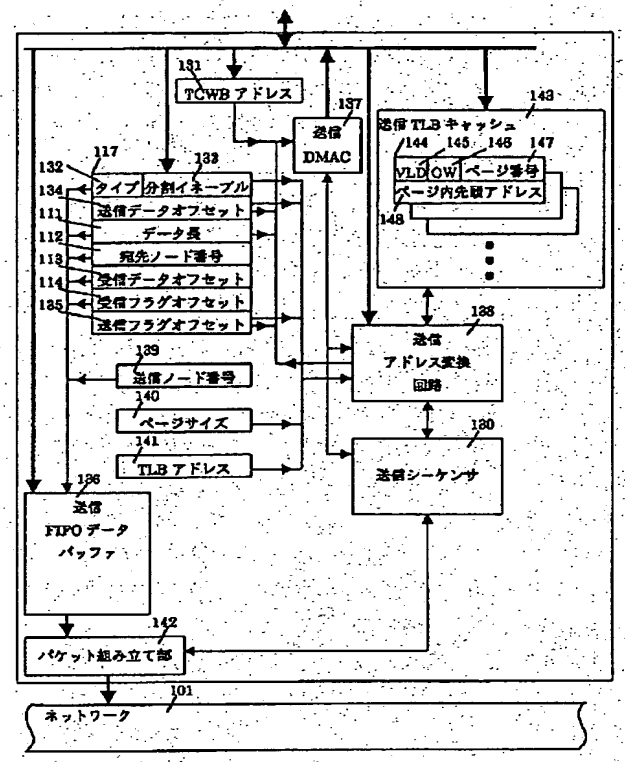
【図１】

図 1



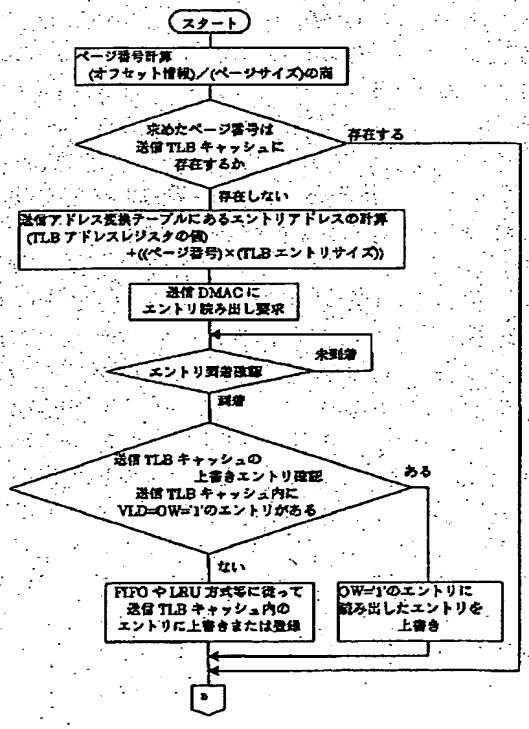
【図２】

図 2



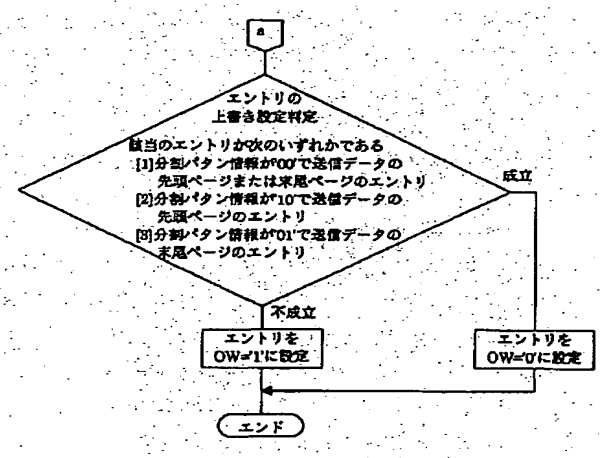
【図３】

図 3



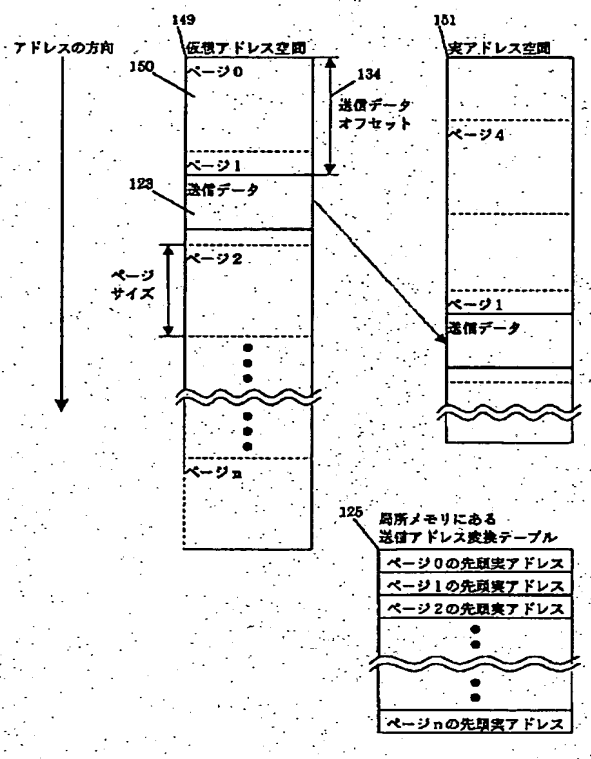
【図４】

図 4



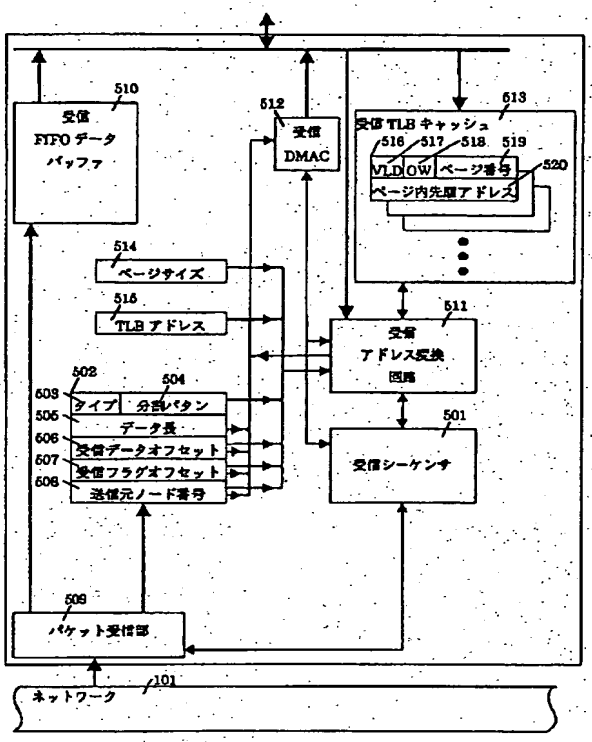
【図５】

図 5



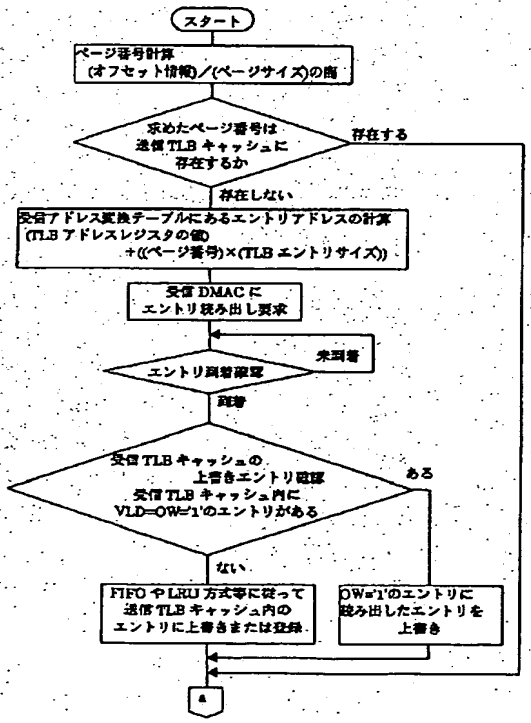
【図６】

図 6



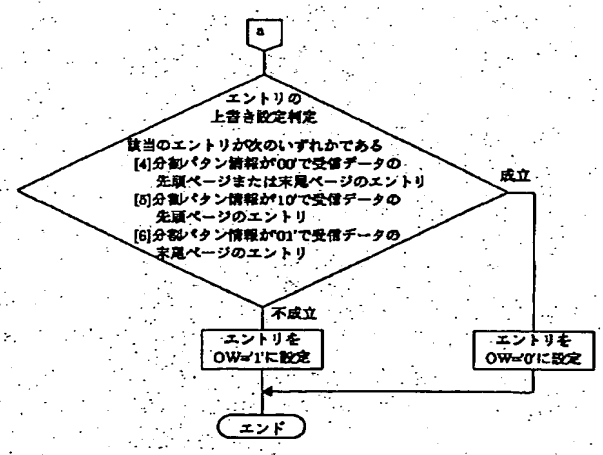
【図７】

図 7



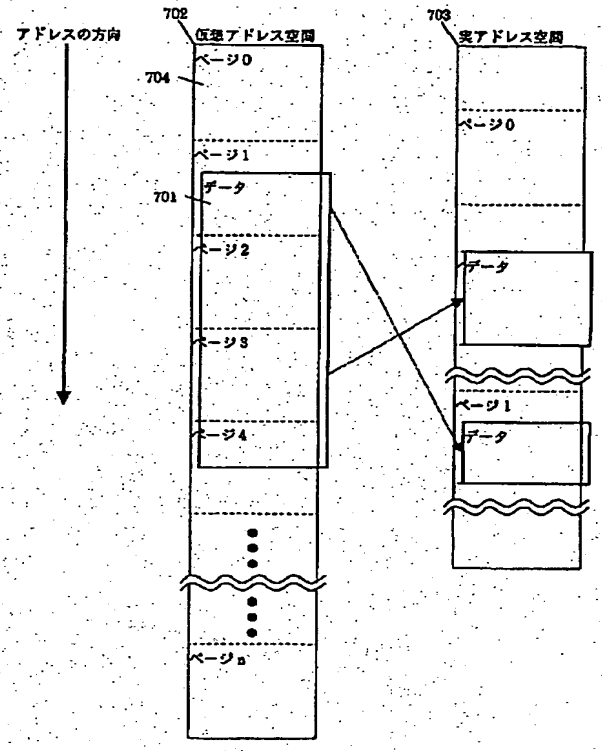
【図８】

図 8



【図９】

図　９



フロントページの続き

(51)Int.Cl.[7]	識別記号	Ｆ Ｉ	ﾃｰﾏｺｰﾄﾞ(参考)
Ｇ０６Ｆ　15/163	６５０	Ｇ０６Ｆ　15/163	６５０Ａ